On the Compatibility of the Use of Algorithms in Parole Decisions and Egalitarian Norms

Emma Søndergaard Jensen

Summative Assignment for PH439 Taught by Dr. Lewis Ross

§

Written in March 2025

Word Count: 3487

Introduction

Parole is a provisional release of a prisoner who has satisfied the minimum requirements of serving his sentence in a prison and agrees to abide by specified behavioral conditions. In the United States, parole is sometimes informed by algorithmic risk assessments that serve as predictive tools, providing a credence about a prisoner's likelihood of reoffending. These decision-making tools, like COMPAS – a U.S. recidivism risk algorithm – have raised concerns about fairness due to their disparate impact on individuals in historically disadvantaged suspect classifications.[1] A suspect classification is a class of individuals who meet a series of criteria suggesting that they are likely to be the subject of discrimination.[2]

In this paper, I argue that the use of algorithms in parole decisions is not unfair and their use is not incompatible with egalitarian norms. As such, the use of such algorithmic systems in parole decision-making does not count as a legitimate form of discrimination.

> **P1.** If the use of algorithms in parole decisions is unfair, then the use of algorithms in parole decisions are incompatible with egalitarian norms.
> **P2.** It is not the case though that the use of algorithms in parole decisions are incompatible with egalitarian norms.
> **C.** The use of algorithms in parole decisions is not unfair.

Section I defends **P1** and **Section II**, **P2**. **Section III** addresses an objection and advances a response.

Section I

P1 stands to be motivated – **I.I**, on COMPAS and preferential algorithms, and **I.II**, on fairness, serve as defense.

I.I. ProPublica's report on COMPAS asserted that the algorithm was "biased against blacks" due to a higher rate of black nonrecidivists incorrectly predicted as 'high risk.'[3] Race – a suspect classification – non-preferentially impacts the COMPAS assessment due to the base rate problem, which disparately impacts (decreases predictive efficacy) those in historically marginalized suspect classes.

[1] COMPAS is the Correctional Officer Management Profiling for Alternative Sanctions program.
[2] In the U.S., these classes are race, national origin, and religion."Suspect Classification." (2024).
[3] Angwin, J. et al. "Machine Bias." (2016).

> *Base Rate Problem* - There is a set B (race categories for parole algorithms): $\{R_W, R_B, R_A, \dots, R_i\}$ where $\forall R_i \in B$ (e.g., R_W is subset 'white'). $\exists x_i$ that instantiates an R_i and x_i will be judged for risk of recidivism. Historically, R_B has a higher observed rate of recidivism in the general population than R_W due to structural inequalities in the criminal justice system. An algorithm trained on set B, refined to calibrate false negatives and positives for recidivism will predict that a given x which instantiates a disadvantaged R ($\neg R_W$), all else equal, will offend at a higher rate. The output of the COMPAS algorithm are such that disparate impact of the false predictions unjustly deprive $x_i \in \neg R_W$ and unjustifiably enrich $x_i \in R_W$ based on general prevalences of the training dataset. Due to unequal overall historical frequencies (base rates), the rate at which $x_i \in \neg R_W$ is unjustly deprived and $x_i \in R_W$ are unjustifiably enriched are the same.

The particular trade-off for COMPAS that inflates its inegalitarian impact is grounded in the base rate problem.[4] COMPAS prioritizes:

- ➢ Statistical Parity - Equal false positive and negative recidivism prediction rates $\forall R_i$,

and not

- ➢ Predictive Parity - Equal ratios of predicted positive to actual positive rates of recidivism $\forall R_i$.[5]

Clearly this trade-off gives rise to disparate impact, and so it is uninteresting to discuss decidedly bad algorithms like COMPAS that clearly violate egalitarian norms. Therefore, for the purposes of this paper, I will be discussing *preferential* algorithms that make a *different* fairness criteria trade-off that aim to ensure equal opportunity through preferential treatment; see **Section II** on **Moral Relevance** for an expanded discussion.[6]

Preferential algorithms are algorithms that artificially construct the competitive situations – here, due parole risk consideration opportunities –that would have been obtained naturally if it were not for the sociohistoric $\neg R_W$ discriminatory practices in the criminal justice system.[7] There is still warranted discussion about whether preferential algorithms for parole decisions violate egalitarian norms because in virtue of their preferential nature, they offer the historically disadvantaged a special corrective process, elevating artificial depression (otherwise caused by privileged base rate consideration) that makes it such that all individuals have just claims to their entitlements. Throughout this paper,

[4] Hedden, B. "On statistical criteria of algorithmic fairness." (2021). p. 211.
[5] Ibid. p. 216 and Angwin, J. et al. "Machine Bias." (2016).
[6] Henceforth, when I employ 'algorithm(s)' let it refer to 'preferential algorithms.' Hedden, B. "On statistical criteria of algorithmic fairness." (2021). p. 214.
[7] McGary, H. "Racism and Justice: The Case for Affirmative Action." (1993). p. 100.

defenses of how preferential treatment is not incompatible fundamentally with the philosophy of egalitarianism will be presented; see **I.II.**, **Moral Relevance**, and **Veracity**. Preferential parole risk algorithms have a fairness-aware architecture that guides that algorithms to use information of group membership to rectify statistical inequities that result from sociohistoric structural injustice.

Egalitarianism treats inequalities as suspect, generally; and many egalitarians have the intuition that inequalities that arise from constitutive luck should be treated differently than those resulting from choices.[8] I suggest that the preferential parole algorithms are corrective for solely the constitutive luck regarding an individual's membership in a suspect B set. In this way, while still propounding a corrective process, the algorithms of focus stay within the bounds of re-constructing the competitive and *equalized* landscape of parole consideration which would have existed but for the disparate treatment of $\forall x_i \in \neg R_W$ and consequent tainted data run-off that gave rise to the base rates problem. Preferential algorithms are compatible with the general philosophy of egalitarianism, then.

I.II. The concept 'fairness' in the fair machine learning (ML) community is "best understood as a placeholder for a variety of normative egalitarian considerations."[9] Operationally, fairness is disciplinarily linked to egalitarianism. Unfairness, as I will consider it, is when an algorithm unjustly deprives an individual (x_i) the equal opportunity for consideration for parole *in virtue of* their membership in a suspect class ($\neg R_W$).[10]

The 'egalitarian norms,' then, can refer to the negation of that unfairness concept. Fairness is tied to x's equal claim to their just entitlement – that of petitioning for parole and being evaluated for it without unjust consideration of or entailment from their membership in $\neg R_W$. To have a preferential parole risk algorithm, then, is to control for x_i's uncontrolled & morally irrelevant membership in $\neg R_W$ (to correct for variations in that luck). If a preferential algorithm is unfair for parole decisions, then it violates egalitarian norms that would otherwise have corrected for constitutive luck that unjustly deprives prisoners of their just entitlement to the fair right to have due consideration for parole.

Fairness criterion in algorithms are how we can assure adherence to egalitarian norms. There is robust literature surrounding which criteria are jointly satisfiable and how fair predictive algorithms should operate. Further, there is agreement in the ML community that adherence to egalitarian norms entail a

[8] Bidadanure, J. and Axelsen, D. "Egalitarianism: Equalizing Luck." (2025).
[9] Binns, R. "Fairness in Machine Learning: Lessons from Political Philosophy." (2018). p. 6.
[10] Hedden, B. "On statistical criteria of algorithmic fairness." (2021). p. 212.

trade-off between fairness criteria because not all plausible and attractive statistical fairness criteria can be met (the *impossibility result* – See **Section III**).[11]

Section II

P2 will be defended through a discussion of preferential algorithms, egalitarian norms, and discrimination through the routes of **Moral Relevance** and **Decision Theory**.

Moral Relevance. "Race neutrality is not attainable."[12] This is the general convergence of ML experts in discussing parole decision algorithms. Optimal trade-offs to remedy historical injustices are, then, the next best thing in building algorithms that adhere to egalitarian norms.[13]

According to James Nickel in "Should Reparations Be to Individuals or to Groups?," preferential treatment is unfair only when derived from morally irrelevant characteristics.[14] The morally irrelevant characteristic that is typically uncorrected for in unfair algorithms is the general historical prevalence of $\neg R_W$ recidivists. Moral relevance in preferential parole algorithms proves compatibility with egalitarian norms by rectifying the unjust deprivation of a due parole consideration caused by statistical fairness criteria trade-offs that disparately impact $\forall x_i \in \neg R_W$.

Say we have the following logical statement: $(\forall x_i)(x_i \leftarrow B \leftrightarrow R_i x_i)$. It roughly translates to if any x_i is a member of B, then x_i is in set B; further, instantiating R_i is then, a property of whatever is in set B. Now let us say that at time t, unjust treatment based on R_i occurred because of the belief R_i was morally relevant. However, at time u, there is a realization that R_i is morally irrelevant. So, the question arises whether it is proper to use x_i's membership in R_i that is always morally irrelevant now that it is time $u + 1$ as morally relevant for the purpose of upholding egalitarian considerations.

Unjust treatment of x_i *in virtue of* R_i membership is considered discriminatory because of the moral irrelevance now of set B. B's moral relevance pre-Civil Rights led to today's disproportionate prevalence of $\neg R_W$ prisoners, which is the sociohistoric origin of the base rates problem. But, to construct a fair parole assessment situation that would have been obtained naturally had it not been for historical discriminatory practices against certain suspect classes, a preferential algorithm *has* to identify R_i as an epistemically relevant feature to correct for in order to serve a preferential and

[11] Ibid. p. 218.
[12] Berk, R. et al. "Fairness in Criminal Justice Risk Assessments: The State of the Art." (2021).
[13] Hedden, B. "On statistical criteria of algorithmic fairness." (2021). p. 211
[14] Nickel, J. "Should Reparations Be to Individuals or to Groups?" (1974). p. 154.

reparative function.[15] Unequal base rates are the foundation of discussions on algorithmic unfairness, and as such, the focus on egalitarian norms advocate for focusing attention on sociohistoric reasons for unequal base rates that affect the predictive efficacy of parole algorithms.[16] Further, because the $\forall x_i \in \neg R_W$ up for parole consideration *prima facie* suffer from statistical disadvantages in a non-corrective algorithm's training, they are morally relevant heirs to the disadvantages of $x_i \in \neg R_W$ at time t. Preferential correction of $\neg R_W$ deprivation is compatible with our **I.II.** definition of fairness because preferentially considering R_i corrects for otherwise unequal parole consideration for x_i in virtue of their R_i membership.

We can use the knowledge of R_i preferentially in parole decisions by aiming for satisfying two fairness criteria. The general recommendation in this paper is to satisfy *Calibration Within Groups (CWG)* fairness criterion lexically prior to predictive parity and aim to fulfill both.[17]

- Hedden's *CWG* - For each possible risk score, the expected percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.[18]

The *CWG* makes it such that the same risk score does not imply different actual risks between groups and predictive parity necessitates reference to individual success in each R_i. *CWP* has been proved a necessary criteria of all 'fair' predictive parole algorithms.[19] If we hold that the relevant egalitarian consideration in parole decision-making is that of ensuring each individual is able to satisfy their claim to just consideration for parole, then correcting for past injustices that make the consideration de-facto unjust without correction is quintessentially egalitarian. This is because the consideration of a morally irrelevant feature R_i at time $u + 1$ rectifies disparate impact on $\neg R_W$s such that R_i becomes epistemically relevant when referred to as a variable contributing to the epistemic value of an algorithm (so that equal predictive parity is satisfied here, in aim). Because $\forall x_i \in \neg R_W$ are more likely to receive a false prediction due to algorithmic oversight that honors *statistical* parity, a corrective feature of a parole algorithm can make it such that R_is are relevant, algorithmically, for ensuring that individuals are not deprived in virtue of belonging to a historically disadvantaged suspect class. $\forall x_i \in \neg R_W$ are *not* more likely to receive a false prediction via an algorithm jointly satisfying *CWG* and predictive parity definitionally.

[15] McGary, H. "Racism and Justice: The Case for Affirmative Action." (1993). p. 100.
[16] Binns, R. "Fairness in Machine Learning: Lessons from Political Philosophy." (2018). p. 8.
[17] Hedden, B. "On statistical criteria of algorithmic fairness." (2021). p. 214.
[18] Ibid.
[19] Ibid. p. 221.

CWG is generally jointly satisfiable with predictive parity because calibration ensures that the proportion of predicted positives/negatives that are legitimate is equal – or, as similar as possible – across groups.[20] Furthermore, the trade-off suggestion of prioritizing *CWG* and then predictive parity over other fairness criteria bypasses the base rate problem because *x*'s recidivism risk is not unjustly inflated due to their membership in $\neg R_W$. Instead, the ratio of predicted positives to real positives (accuracy rate) is calibrated to be equal and is concordant in meaning between $\neg R_W$ and R_W. Calibration, with the preferential algorithm, will pertain to correcting the larger ratio in $\neg R_W$ through adversarial ML training. That is, insofar as the preferential algorithm adheres to predictive parity and *CWG*, it will use information regarding statistics of R_i in a way that is corrective for sociohistoric injustice. Moreover, no prisoner is deprived the equal opportunity for consideration for parole *in virtue of* their membership in a suspect class. Our preferentially-employed fairness criteria do not deprive any given *x* in virtue of their membership (even $x_i \in R_W$) and only correct for calibrated rates of *x*'s whose opportunity for consideration is unequal due to structural injustice that disables satisfaction of the necessary fairness criterion (*CWG*). Framed as a contextually informed calibration, preferential parole algorithms are compatible with egalitarian norms and because of the moral relevance of sociohistoric deprivation of $\forall x_i \in \neg R_W$ renders the use of R_i justified and not a legitimate form of discrimination.

Decision Theory. Through decision theoretic considerations, we can see that algorithmic parole decision-making – preferential *or not* – *can* be compatible with egalitarian norms. The compatibility depends on two considerations: supplementary human reference to the algorithm and the invariance feature of partition independent algorithms.

First, algorithmic decision-making provides a consistent methodology.[21] Human decision-making is prone to faltering in the wind of unforeseen bias whereas algorithms have predictable biases. Known algorithmic biases observe a limit such that they do not dominate human psychology. Humans foresee demonstrated consistency and can make a decision about parole in tandem with a parole board interview in order to make a more informed decision about whether a prisoner should be granted parole. Additionally, the inegalitarian consequences of general (non preferential) algorithms can be mitigated on the front end by testing and refining the algorithm.[22] Knowing the consistent biases the algorithm turns out by refining the credence weighting that results in the probability distribution that

[20] Chouldechova, A. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Instruments." (2017).
[21] Sunstein, C. "Governing By Algorithm? No Noise and (Potentially) Less Bias." p. 3. (2021).
[22] Ibid. p. 4.

informs the final assessment, one can train the machine to observe finer results. This, more generally, is explicit convergence with *CWG*.

Further, in cases of risk assessment for reoffense and parole decisions, the decision-makers are operating in a case of uncertainty. Here, the action decision-makers can take is:

- ➢ Well-Defined - They can reach a decision about whether to release the prisoner on parole or not),
- ➢ Somewhat Modellable - They have access to algorithms and processes that can give them an idea about future success,
- ➢ Grounded with a General Basis in Probability - Comparison with previous cases where certain features were present and outcomes correlated.

In these situations, the most effective way to reach a rational decision is through probability-based risk analysis, consulting multiple views and ranking those assessments.[23] The use of algorithms in parole decisions does not negate the possibility for their consultation in tandem with parole interviews; after all, they are not designed to be used in a void.

Algorithms in parole decision-making also depend on probabilities – especially those with preferential aims. This leads to a discussion of the invariant feature of partition independent algorithms in parole decision-making. Say, in some amount of cases where a prisoner will not reoffend, details φ, ψ, γ, and α were present at varying rates. One can use those correlations to estimate a quantitative level for the descriptive burden of what qualifies as a potentially successful parole. Partition independent algorithms yield consistent – invariant – risk probabilities, regardless of how states of the world are categorized. The invariance that comes from partition independence promotes equality of claim because regardless of the partitioned states of the world (features), the probability distribution that informs the risk assessment remains consistent regardless of how other prisoners are categorized. So, how would this work?

Assign credences to the details – φ, ψ, γ, and α – in successful parole cases which will likely mirror their prevalence. Have, then, these details be descriptive of data that is preferential and choice-based: φ could be the number of communities formed in prison, ψ could be whether they understand that the crime they committed was wrong, γ could be the original offense type, and α, the number of previous offenses. While these details vary in explanatory nature, they are descriptive facts of individual cases.

[23] Speigelhalter, D. "The Art of Uncertainty: Living with Chance, Ignorance, Risk & Luck." (2025).

Descriptively accurate partitions are not bad nor unfair – they are simply comprehensive.[24] Invariant predictive efficacy ensures that a prisoner's risk assessment is based on the probability distribution of their own characteristics and not how their group is defined. This side steps our unfairness concern (**I.II.**). By ensuring that the reoffense risk probability distribution that the parole decision algorithm outputs is partition independent, it promotes individual-level fairness, ensuring that it is not possible to form strategic group boundaries that produce unjust outcomes for $\forall x_i \in \neg R_W$.

Section III

Literature has demonstrated that a number of intuitively attractive statistical criteria of fairness are not jointly satisfiable.[25] This is known as the *impossibility result*. Critics of **P2** argue, concerned that the impossibility results necessitates unfair algorithms, that parole algorithms are incompatible with egalitarian norms. They claim preferential algorithms cannot provide fairness due to deliberate disparate programming based on protected characteristics and mathematical challenges that entail algorithmic unfairness.[26] These critics raise important points, however they are mistaken about the first's relevance and the second's veracity.

Relevance. In the U.S., where our discussion has been centered, suspect classifications are subject to 'strict scrutiny' for fair and just use. These protected characteristics can only be deliberately used when it is reasonably tailored and fulfills a compelling government interest.[27] I argue that use of partitioning (via including invariant preferential corollaries to R_is) preferential parole algorithms fulfill both components of strict scrutiny. Further, the permit of assessing a prisoner who has served minimum time and their just claim to parole is indeed a compelling government interest. The fair assessment and subsequent granting of conditional liberty that is properly based in a predictive effectiveness that is descriptively accurate is a right. Failing to promote equality in satisfying that right violates foundational government commitments. Preferentially partitioning in this way is also narrowly tailored because it would be unreasonable to substitute another classification due to being compelled to sidestep the base rate problem which is fundamentally rooted in the classification. There is simply no race-neutral way to progress past the mathematical challenges that different demographic prevalence in prisons resulting from historical injustices present in parole decision-making.[28] And, at least if one constructs, tests, and revises an algorithm that preferentially uses these factors (and nothing is to say

[24] Sunstein, C. "Governing By Algorithm? No Noise and (Potentially) Less Bias." p. 5. (2021).
[25] Hedden, B. "On statistical criteria of algorithmic fairness." (2021). p. 216.
[26] Binns, R. "Fairness in Machine Learning: Lessons from Political Philosophy." (2018). p. 8.
[27] "Strict Scrutiny." Cornell Law School. (2024).
[28] Berk, R. et al. "Fairness in Criminal Justice Risk Assessments: The State of the Art." (2021).

that the algorithmic coding is proprietary), then decision makers can know what is informing the consistent, invariant prediction.

Veracity. I argue that the *impossibility* result is overinterpreted; we, optimistically, ought not think that this means all predictive algorithms are unfair.[29] Not all statistical criteria are necessary conditions for an algorithm to be fair; in fact, only *CWP* is.[30] The violation of various fairness criteria does not entail an algorithm's unfairness because fairness in ML is about trade-offs between fairness criteria and finding those that are compatible with one another while also bypassing the base rate problem (the foundational egalitarian concern of this discussion), which *CWP* and predictive parity together accomplish. Additionally, suppose it is true that joint satisfaction of *CWP* and predictive parity turns out less accurate predictions. My argument still does not depend on this preferential algorithm's *accuracy*. Rather, my discussion is concerned with *egalitarian norms*. Therefore, if the combination of fairness criteria I suggest is not especially conducive to elevating predictive accuracy but the criteria are still compatible with each other and are together creating an algorithm that fulfills egalitarian norms and does not discriminate, then my argument still succeeds.

Conclusion

Trade-offs that prioritize egalitarian norms are necessary for preferential algorithms – those that do not deprive x_i's just consideration for parole *in virtue of* their $\neg R_W$ membership. Preferential algorithms are not necessarily incompatible with egalitarian norms because there exists jointly satisfiable fairness criteria (which fulfills the necessity requirement via *CBG*) that avoids the base rate problem and ensures that both groups – calibrated via R_is – observe predictive parity. While the *impossibility result* remains an important concern in the bigger picture of fairness in predictive parole algorithms, it is less concerning as a result that impacts the *actual* egalitarian norm satisfaction of an algorithm and more so, interesting as a concept of how fairness and egalitarian norms are built into algorithms. It is held, then, that the use of algorithms in parole decisions is not unfair and their use is not incompatible with egalitarian norms. Further, the use of preferential algorithmic systems in parole decision-making does not count as a legitimate form of discrimination due to how the information regarding R_is is used (e.g., it does not deprive $\forall x_i$ unjustly of a due parole consideration *in virtue of* $x_i \in R_i$).

[29] Hedden, B. "On statistical criteria of algorithmic fairness." (2021). p. 211.
[30] Ibid. p. 221.

References

Angwin, J. et al. “Machine Bias.” *ProPublica*. (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Berk, R. et al. “Fairness in Criminal Justice Risk Assessments: The State of the Art,” Sociological Methods and Research 50, no. 1: 3–44. (2021).

Bidadanure, J. and Axelsen, D. “Egalitarianism: Equalizing Luck.” *Stanford Encyclopedia of Philosophy*. (2025).

Binns, R. “Fairness in Machine Learning: Lessons from Political Philosophy.” *Conference on Fairness, Accountability, and Transparency,* 81: 1 - 11. (2018).

Chouldechova, A. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Instruments.” *Liebert Publising*. (2017).

Hedden, B. “On statistical criteria of algorithmic fairness.” *Philosophy and Public Affairs*. p. 209 - 231. (2021).

McGary, H. “Racism and Justice: The Case for Affirmative Action.” *Science and Society* 57: 99 - 100. (1993).

Nickel, J. “Should Reparations Be to Individuals or to Groups?” *Analysis* Vol. 34, No. 5: 154 - 160. (1974). https://doi.org/10.2307/3327632.

Speigelhalter, D. “The Art of Uncertainty: Living with Chance, Ignorance, Risk & Luck.” *Lecture Series at the London School of Economics*. (2025).

“Strict Scrutiny.” *Legal Information Institute*. Cornell Law School. (2024). https://www.law.cornell.edu/wex/strict_scrutiny.

“Suspect Classification.” *Legal Information Institute*. Cornell Law School. (2024). https://www.law.cornell.edu/wex/suspect_classification.

Sunstein, C. “Governing By Algorithm? No Noise and (Potentially) Less Bias.” *Duke Law Journal*. p. 1 - 17. (2021).

London School of Economics, UK. (2025).